Evidence in Classical Statistics

Samuel C. Fletcher[*] and Conor Mayo-Wilson[†]

In 2012, CERN scientists announced the discovery of the Higgs boson, claiming their experimental results finally achieved the 5σ criterion for statistical significance. Although particle physicists apply especially stringent standards for statistical significance, their use of "classical" (rather than Bayesian) statistics is not unusual at all. Classical hypothesis testing—a hybrid of techniques developed by Fisher, Neyman and Pearson—remains the dominant form of statistical analysis, and p-values and statistical power are often used to quantify evidential strength.[1]

The dominance of classical statistics raises a puzzle for epistemologists. On one hand, science is a paradigmatic source of good evidence, with quantitative experimental science often described in classical statistical terms. On the other, the hybrid of Fisherian and Neyman-Pearsonian techniques is generally rejected by philosophers, statisticians, and scientists who study the foundations of statistics.[2] So why is the use of classical statistics in empirical science so epistemically successful? Do classical "measures" of evidence actually measure anything epistemically important?

This chapter provides some positive answers to these questions. Section 1 quickly reviews classical hypothesis testing, including the Fisherian, Neyman-Pearsonian, and hybrid versions. We focus on hypothesis testing rather than estimation (of the value of a parameter) because there are closer parallels between the former and philosophical theories of evidence, justification, and knowledge.[3] The logic for evidence it proposes, however, is only partial and consequently so are its ambitions to justify belief and constrain

[*]Department of Philosophy, University of Minnesota. Email: scfletch@umn.edu

[†]Department of Philosophy, University of Washington. Email: conormw@uw.edu

[1] See section 3.5 of Gigerenzer et al. [1989] for a history.

[2] See chapter five of Howson and Urbach [2005]. Even proponents of classical methods such as Mayo [1996, 2018] argue that the hybrid approach needs additional concepts (e.g., severity).

[3] A reduction of estimation to hypothesis testing would draw the parallels closer, but we do not attempt to take a stand on the relationship between estimation and testing here.

rationality. Yet it does seem to meet more fully the epistemological goals of indicating truth and guiding inquiry objectively and publicly [Kelly, 2016].

In contrast with epistemologists' concerns, though, most traditional questions about the *ontology* of evidence—e.g., is evidence propositional? Is it sense data, mental states, or something else?—are less apt in discussions of classical statistical evidence, which, beyond just assuming that evidence consists somehow in recorded data, is largely quietist about ontology. Part of the reason for this is that that classical statisticians typically understand this data as relatively stable and mind-independent. Perception, memory, and computational constraints, therefore, are typically ignored in theories of statistical evidence.[4]

Regardless, the discussion in this first section reveals how evidence in classical statistics bears similarities to tracking and anti-luck approaches to knowledge and justification. Section 2 reviews objections to naive uses of classical statistics as analogous to well-known criticisms of simplistic tracking theories of knowledge.[5] We then illustrate several ways in which recent developments in hybrid classical testing (e.g., by Mayo and Spanos [2006, 2011]) are analogous to modal principles (e.g., basis-relative versions of sensitivity) that are used in more sophisticated tracking theories of knowledge. In particular, comparing classical statistics with contemporary tracking theories allows us gain insight into the scientific and quantitative application of reliabilist epistemology.[6] This suggests a possible reliabilist explanation of why hitherto unrectified classical hypothesis testing, despite its numerous problems, has been so successful in practice.

Classical statistics remains a largely unmined source of examples and concrete applications for epistemological theories of evidence, justification, and the like. Our concluding section 3 thus outlines various directions for research concerning the relationship between classical statistics and philosophical theories of evidence.

1 Classical Hypothesis Testing

To illustrate different forms of classical hypothesis testing in what follows, we will rely on an example. Imagine Ada is interested in whether, at public

[4] Theories of justification for logical and mathematical beliefs often do the same.

[5] For similar criticisms, see Mayo-Wilson [2018], who argues that just as some tracking and relevant alternative theories violate epistemic closure in unacceptable ways, naive use of classical methodology requires one to endorse claims like, "The data provides evidence that smoking causes lung cancer, but it is not evidence that smoking causes some disease."

[6] See [Fletcher] for a detailed theory.

universities in the United States, male and female assistant professors were offered different starting salaries, on average, between 2000–2010—possibly also in what gender equity policy to recommend to administrators based on her findings. Using publicly available data, she randomly selects 30 men who were hired as assistant professors during this period and 30 women. She calculates an average salary of $\$68,326$ for men and $\$67,552$ for women. How does this data provide evidence about different answers to her question, and what recommendations she should offer?

Scientific (and pragmatic) questions such as these are qualitative, invoking no probabilistic claims. By contrast, classical hypothesis testing demands that different answers—*hypotheses*, in statisticians' jargon—must each fix a probability distribution for the data gathered. Whatever assumptions fix this distribution are collectively called a *simple statistical hypothesis*, while a *composite statistical hypothesis* is a (perhaps infinite) disjunction of mutually exclusive simple ones. Formally, if $\mathcal{F}$ denotes the possible data collected and $\mathbb{P}(\mathcal{F})$ all probability measures on $\mathcal{F}$, then one can represent a simple statistical hypothesis as element of $\mathbb{P}(\mathcal{F})$, and a composite one by a subset of $\mathbb{P}(\mathcal{F})$.

Now, sometimes an answer to a scientific question on which the possible data $\mathcal{F}$ bear—a *scientific hypothesis*, in some statisticians' jargon—is quantitative enough to entail a simple statistical hypothesis $P \in \mathbb{P}(\mathcal{F})$. In the case at hand, however, it is not: to test any hypothesis about the average gender differences in salaries, Ada must invoke and justify auxiliary assumptions about the probability of selecting at random 30 different salaries each for men and women. For example, Ada may reasonably assume that each of her 30 samples was, for each gender, drawn truly at random and independently of each other, with each salary in the database equally likely to be drawn (for each gender). The exact contents of the database will then determine P for salaries of both male and female assistant professors, respectively.

However, the reason Ada is sampling from the database is that she does not feasibly have access to its whole contents. If she did, she would only need to engage in *descriptive* statistics, summaries of the database contents themselves. Indeed, lack of knowledge about the *population* of interest is typically one of the main motivations for the development of *inferential* statistics, such as statistical testing. Thus she will need to make some justified auxiliary assumptions about general features of P—that is, to form a well-circumscribed composite statistical hypothesis—and let the details be specified by a choice of *parameter*, a set of numbers each of which determines a simple statistical hypothesis.[7] For example, Ada might assume that whatever the distribution of salaries in the database, the probability distribution for the mean of a 30-element sample from it is (approximately) normal, with the mean of that normal distribution depending on the gender sampled.[8] This would allow her to consider simple statistical hypotheses, such as that the means of the two gender's distributions were equal to a specific value, e.g., the same value. It also allows her to consider composite statistical hypotheses, such as that the means of the two gender's distributions differ by at most some number ϵ which (for instance) she considers to be practically significant. Researchers typically call the hypothesis under test that test's *null hypothesis.* How such a test bears on a null hypothesis depends on the type of testing used.

1.1 Fisherian Testing

Fisherian testing quantifies purely the evidence *against* the null hypothesis. Because, in this sense, it provides only negative evidence, it facilitates a probabilistic sort of falsificationist reasoning: the stronger the evidence against the null hypothesis, the more reason one has to reject it, although Fisherian testing itself does not provide complete guidance on rejection. (We'll return to a related framework for testing that provides more such guidance in section 1.2.)

The fundamental concept of Fisherian testing, its technique for quantifying negative evidence, is the *p-value.* Intuitively, the p-value associated with a statistic of the recorded data is the probability of observing data at least as extreme—i.e., improbable—as that actually recorded, if the null hypothesis were true. Small p-values thus indicate the data is strong evidence against the null, as a p-value will be small only if the observed data, and any data at least as extreme than it, were unlikely to have been observed if the null hypothesis were true. However, a small p-value does *not* generally indicate evidence *for* any specific simple alternative statistical hypothesis, and a large p-values does *not* indicate evidence for the null hypothesis,for

[7] Statisticians make a distinction between *parametric* inferential statistics and *non-parametric* statistical inference, but these names are a bit misleading: the former concerns families of probability distributions indexed by a finite set of numbers, while the latter allows this set to be infinite [Tsybakov, 2009].

[8] The assumption of a normal distribution is an *approximation*, as it entails positive (albeit small) probabilities for data that is impossible, such as negative salaries, or salaries whose magnitude is larger than the world's yearly economic output. Although the use of approximation and idealization in statistical testing is important and pervasive, it engenders much the same issues as it does elsewhere in science.

reasons we'll discuss in section 2. In any case, p-values depend on three components: (i) the *actual/observed* data, (ii) an "extremeness" relation on *possible* data sets related to their probability, and (iii) the null hypothesis, which determines these probabilities. So the fundamental measure of evidence for Fisherian testing depends, in a crucial way, upon modal facts about *possible data*.

To illustrate a formal definition for p-values,[9] consider again the case of Ada's investigation of new faculty salaries, and suppose that she is interested in performing a Fisherian test of the simple statistical hypothesis that there is no difference in mean starting salaries in the United States between 2000–2010 among male and female assistant professors. (Recall that the Fisherian test allows one to find evidence against this hypothesis, but not for it nor for any other simple statistical hypothesis.) Suppose that $D \in \mathcal{F}$ represents the data Ada observed and that her simple null hypothesis determines a probability distribution $P \in \mathbb{P}(\mathcal{F})$. Further suppose that the extremeness relation on possible data sets is given by their probability ordering by P—formally, for any $E_1, E_2 \in \mathcal{F}$, let $E_1 \leq_P E_2$ abbreviate the claim that $P(E_1) \geq P(E_2)$.[10] Then Ada's p-value equals $P(\{X \geq_P D\})$, where $\{X \geq_P D\} = \{E \in \mathcal{F} : E \geq_P D\}$ is the set of possible data that are at least as extreme as what Ada actually observed. Since the observed difference between the average salaries of male and female assistant professors was \$774, this amounts to $P(\{|X| \geq_P \$774\})$. Whether the p-value is large or small will depend as well on the variance of the normal distributions hypothesized.

[9] Despite widespread agreement about how to calculate p-values in most cases, there is surprising lack of agreement on the formal definition of p-values. Instead of defining a p-value in terms of an "extremeness" relation, Wasserman [2004, p. 157] defines it to be function of a collection of *statistical tests* (as well as the data and null hypothesis). Lehmann and Romano [2010] do the same, except their definition of p-value is applicable only when the statistical tests have a particular property. The definition of Casella and Berger [2001, p. 397] is completely different; they define a p-value to be a statistic with a particular type of distribution. The definitions, moreover, are not mathematically equivalent to each other, or to the one above in terms of an extremeness relation. This makes it unclear how to apply Fisherian testing in some novel contexts.

[10] If the null hypothesis specifies a continuous distribution, then the extremeness relation should be specified using the probability density, not the probability itself. Whatever the distribution specified, though, one more generally lets the extremeness relation be defined in this way through the probability (density) of a function (statistic) of the data. How to choose such a function is an interesting topic on which Neyman-Pearson (section 1.2) and hybrid (section 1.3) approaches to testing have much to recommend [Casella and Berger, 2001, Ch. 8.3.2], although the seeming latitude in this choice does raise questions about the objectivity of classical statistical testing.

As we have illustrated, p-values are typically calculated for simple statistical hypotheses. When the null hypothesis is composite, researchers typically define its p-value to be the maximum (or supremum) of the p-values of the disjuncts. This definition, which is rarely explained, is similar to worst-case reasoning. Unless one's data is improbable *for every way* that a composite null hypothesis might be true, one's data will not yield a low p-value and will not count as evidence against the hypothesis according to Fisherian testing.

1.2 Neyman-Pearson Testing

Fisher aimed to quantify the strength of evidence that a *particular* data set provides against a *particular* null hypothesis. In their landmark 1933 paper, Neyman and Pearson (NP) abandon that goal completely for statistical testing. They write:

> We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.
>
> But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour . . . [that] insure that, in the long run of experience, we shall not be too often wrong . . . Such a rule tells us nothing as to whether in a particular case [a hypothesis] H is true . . . But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false. [Neyman and Pearson, 1933, pp. 141–142]

The fundamental concern of NP's theory is with the long-run reliability of statistical testing *procedures* to minimize (what contemporary philosophers would call) epistemic risk. To do so, they must consider not just the (simple null) hypothesis under test, but also an alternative, simple or composite. For NP, a statistical testing procedure is simply a rule that partitions $D \in \mathcal{F}$ into two sets, A and R: those that warrant accepting the null hypothesis (and rejecting the alternative), and those that warrant accepting the alternative hypothesis (and rejecting the null), respectively. There are, of course, in general many such partitions available, so testing procedures ought to be judged by how well they minimize two types of errors. Let Θ denote the set of simple statistical hypotheses composed by the disjunction of the null hypothesis $\theta_0 \in \Theta$ and the alternative hypothesis (which, again, may be composite). Each $\theta \in \Theta$ prescribes a distinct probability distribution $P_\theta \in \mathbb{P}(\mathcal{F})$. Then, a Type I error occurs when θ_0 is true—i.e., the probability of the data is actually given according to P_{θ_0}—but $D \in R$. A Type II error occurs when θ_0 is false—i.e., the probability of the data is actually given according to P_θ for some $\theta \neq \theta_0$—but $D \in A$. The *significance level* or *size* α of a testing procedure is the probability of a Type I error when θ_0 is true: $\alpha = P_{\theta_0}(\{D \in R\})$.[11] The testing procedure's *power* at $\theta \in \Theta$, $\beta(\theta)$, is one minus the probability of a Type II error when θ is true: $\beta(\theta) = 1 - P_\theta(\{D \in A\})$. Notably, the size of a testing procedure is a constant (assuming, as is typical, that the null hypothesis is a simple statistical hypothesis), while its power is a *function* of the simple statistical hypotheses comprising the alternative. It's generally not possible to find a testing procedure that minimizes both the Type I and Type II error rates at once. The tradition in NP testing is to fix the size of the testing procedures of interest conventionally, and then maximize the power, although there are good decision-theoretic reasons to balance them together rather than sequentially [Cox, 1958].

Details about how to choose a testing procedure aside, already there are two important points to note about NP testing in general. First, it assumes the truth of the disjunction of the null and alternative hypotheses, even when this disjunction is not a tautology. This is clearly also an idealization in much the same sense as discussed in footnote 8; whether it is a warranted idealization must be judged from other contextual knowledge and evidence. Second, size and power are attributes of testing procedures, not of particular tests applied to particular data. These two quantities, therefore, can be calculated for one's anticipated testing procedure before any data is gathered whatsoever. Their lack of dependence on the data undergird NP's denial that the size and power of a test quantify any evidence that a particular data set provides for any particular hypothesis. Yet if one believes and acts according to the verdicts of NP tests of sufficient size (about 1%, as NP suggest), then the long-run reliability of those testing procedures is supposed to provide evidence that *most* of one's beliefs are true, even if they do not

[11] Some authors distinguish the size of a testing procedure, as we have defined it, from the significance level, by which they denote only a strict upper bound on the Type I error [Casella and Berger, 2001, p. 385]. The relevance of this distinction becomes important only for more complicated testing procedures, the details of which are beyond the present scope of discussion.

provide any evidence for the truth of any particular belief (associated with a simple statistical hypothesis).

1.3 Hybrid Approaches to Testing

Fisherian and NP testing differ in at least two important ways. First, they relate different sorts of relata. While both concern (simple) null hypotheses and data, only NP essentially considers particular alternative hypotheses through its use of the power concept. The significance of this additional relata is that it is supposed to provide warrant for positive evidence, i.e., acceptance of hypotheses, while Fisherian testing generally provides only negative evidence. Indeed: "It is sometimes suggested that how confident a scientist is justified in being that a given hypothesis is true depends, not only on the character of relevant data to which she has been exposed, but also on the space of alternative hypotheses of which she is aware" [Kelly, 2016]. Similar suggestions have been made about statistical evidence [Earman, 1992].

However, the way that NP testing implements this relation has been subject to withering criticism. In particular, its long-run behavioristic orientation, advocated by Neyman in particular, supports no evidence relations bearing on any particular hypotheses, only a probabilistic guarantee that one's actions, based on the hypotheses "accepted" through NP testing, would often be apt [Howson and Urbach, 2005]. Fisher thus called Neyman's philosophy "childish" and "horrifying [for] intellectual freedom in the west" [Gigerenzer et al., 1989, pp. 106–109]. This proscription of evidence relations for particular hypotheses was extreme enough that Pearson himself later rejected it [Mayo, 1996, Ch. 11].

Yet, Fisherian testing is not an entirely satisfactory alternative, for it provides no way for data to *support* an hypothesis. Consequently, both Fisherian and NP testing are often taught side-by-side in contemporary statistics textbooks, without mention that the two sets of techniques are based in radically different conceptions of the relation between data and hypotheses. There's often a vague sense, in these hybrid accounts, in which evidence for an alternative hypothesis requires both a low p-value and sufficiently high power, even though the former is a function of the data whereas the latter is a property of the testing procedure only. If the power of a testing procedure warrants only the reliability of certain behavior in the long-run, how is it to help warrant evidence for hypotheses in particular cases?

Perhaps surprisingly, statisticians have not directly addressed these issues.[12] But philosopher Deborah G. Mayo [Mayo, 1996, 2018] has, often in collaboration with econometrician Aris Spanos [Mayo and Spanos, 2006, 2011]. To understand their suggestion, which they call *severe testing*, it is first helpful to understand the relationship between p-values and significance levels. It often the case that one motivates the selection of the rejection region $R \subset \mathcal{F}$ of an NP test of a simple hypothesis θ_0 by appeal to possible data that would yield low p-values. Explicitly $R = \{D \in \mathcal{F} : P_{\theta_0}(X \geq_{P_{\theta_0}} D) \leq \alpha\}$, where α is the significance level. Comparing with the definition of the p-value provided above, this amounts to letting the rejection region be those data sets that yield a p-value no larger than the chosen significance level, and the acceptance regions be those that yield a p-value larger than the chosen significance level. According to this prescription, the significance level and the p-value are calculated in a similar way, except the former, applied as it is to a testing procedure, does not depend on any particular data collected, while the former does so depend, and therefore quantifies the fit or compatibility of that particular data with the null hypothesis.

Mayo and Spanos suggest introducing a quantity, what they call the *severity* of a particular test, analogous to power as p-values are to significance levels. Severity is, in other words, a data-dependent version of power, and its application should help underwrite warrant for positive evidence for an hypothesis. The severity with which θ_0 is tested against θ by data D is given by

$$SEV(D, \theta) = P_\theta(X >_{P_{\theta_0}} D) = 1 - P_\theta(X \leq_{P_{\theta_0}} D).$$

When the alternative hypothesis is composite, they propose calculating the supremum of this function over each simple θ comprising the composite alternative.[13] In a word, a hypothesis H_0 is severely tested—receives a high severity score—just to the extent that it's likely that more extreme data (according to H_0) would have been observed if the alternative hypothesis were true. A test of θ_0 with data D then produces evidence for θ_0 to the degree that both the p-value and the severity are high.

Fisherian, NP, and hybrid approaches such as Mayo and Spanos's severe testing share a commitment to the dependence of evidence upon $\mathcal{F}$, the

[12] There is a literature on so-called post-hoc/restrospective analyses of power, but these techniques do not accomplish their goals [Hoenig and Heisey, 2001].

[13] In section 2 we discuss parallels between severe testing and modal criteria for knowledge; this particular proposal for severity when the alternative hypothesis is composite is a slight departure from these parallels because it does not consider only hypotheses sufficiently similar to the null, as modal conditions considering only the closest possible worlds do. This might provide grounds for modifying severe testing [Fletcher].

space of possible data, not just the actual data. Most Bayesian statisticians reject this claim. For example, Lindley [1971, p. 436] writes that "unbiased estimates, . . . sampling distributions, significance levels, power, all depend on something more—something that is irrelevant in Bayesian inference—namely the sample space." If Bayesian inferences depend only on one's posterior probability distribution over hypotheses, then this is determined by one's prior probability distribution over hypotheses and the likelihoods of the data, but not essentially on one's prior probability distribution over the possible data generally. Moreover, these approaches also require an ordering of "extremeness" on these possible data sets. This structure on possibilia should be reminiscent to epistemologists of the closeness, or similarity, of possible worlds invoked in reliabilist theories. We indeed turn now to the relationship between modality and evidence in classical statistics in section 2.

2 Reliablism, Modality, and Classical Statistics

There are important parallels between the theories of classical hypothesis testing outlined in section 1 and reliabilist (e.g., tracking and anti-luck) theories of knowledge. In the former, the evidence for an hypothesis from data depends not just on that hypothesis and the actual data, but also (1) the merely *possible* data, and, in the case of Neyman-Pearson and hybrid approaches, (2) the possible alternative hypotheses. Reliabilist theories of knowledge regiment which beliefs count as knowledge through modal conditions, which may depend upon (1) the *basis* of a belief (i.e., observations and the method for forming the belief) and (2) the reasons the belief might have been false. Conditions such as adherence and sensitivity employed in these theories parallel the size and power concepts in classical statistical testing. Thus, criticisms of these conditions, which motivate their basis-relative versions, find analogs in statistical criticisms of size and power, which lead to data-dependent versions thereof in the hybrid approaches of Mayo [1996, 2018] and Mayo and Spanos [2006, 2011].

Stine [2008], Dretske [1971], and Nozick [1981], among others, famously argue that, if a subject S knows some proposition φ, then her belief in φ must be *sensitive*, i.e., if φ had been false, then S would not have believed φ.[14] Although sensitivity captures many intuitions about knowledge, many, including Nozick himself, argue the crude formulation just given is not nec-

[14] [Becker and Black, 2012] contains several more recent articles on the plausibility of sensitivity.

essary for knowledge.

Nozick [1981] asks us to imagine an elderly grandmother who is dying in a hospital. Her grandson, a professional stuntman, visits her in the hospital after a near-death incident on a movie set. But suppose that, if the stuntman had died, his grandmother would have remained blissfully ignorant until her death. Intuitively, the grandmother knows that her grandson is alive after seeing him, but by stipulation, her belief is not sensitive: she would have falsely believed her grandson to be alive even if he had died. Sensitivity, Nozick concludes, is not necessary for knowledge.

Power, as defined in section 1.2, is the statistical analog of sensitivity: a classical statistical test is sensitive to an alternative hypothesis to the extent that it will probably reject the null hypothesis when that alternative is true. So unsurprisingly, high power has been criticized as necessary for statistical evidence using cases analogous to Nozick's grandmother example. To see why, we modify an example due to Berger and Wolpert [1988, pp. 5–6, Example 1].

Example 1 Imagine Acme Corp. makes two types of urns, both of which contain 100 balls. R-urns contain 1 ruby ball, 50 scarlet balls, and 49 black balls; B-urns contain 49 scarlet balls and 51 black balls.

Urn Type	Ruby	Scarlet	Black
R	1	50	49
B	0	49	51

Table 1: Colors of the balls in the Acme urns.

You find an Acme urn, but you don't know which type. So you decide to run a Neyman-Pearson test on the hypothesis H_R, "The urn is R-Type." You will draw one ball from the urn. If you draw a black ball, you'll reject H_R for H_B, "The urn is B-Type." Otherwise, you'll accept it.

Your test has low power: there's a 49% chance you will accept H_R if H_B is true. Yet, observing a ruby ball entails H_R, hence is conclusive evidence for H_R and against H_B. So, having high power cannot be a necessary condition for a test to yield strong evidence.

The problem with power, and with sensitivity, is that they depend only on procedural aspects of gathering evidence, not on the particular details of the evidence gathered. Once one recognizes the analogy between sensitivity and power, it's unsurprising that proponents of each have modified their theories in analogous ways to avoid the above counterexamples. For example, Nozick notices that, if the stuntman had died, his grandmother would not have believed him to be alive *on the same basis/evidence*. That is, if her grandson had died, the grandmother's belief would have been sustained by different evidence altogether, namely, that the last time she saw her grandson, he was alive and well. So, roughly, Nozick argues that, if S knows that φ, then S's belief need be sensitive only in the sense that if φ had been false, then S would not believe φ on the same basis.[15] And similarly, Mayo [2018] and Mayo and Spanos [2006, 2011] emphasize that regardless of the power of a test, in order for it to provide evidence for a null hypothesis against an alternative, the data actually collected must yield high severity (as defined in section 1.3): if the alternative hypothesis were true, the there would be a high probability that the data collected would have been more extreme than they actually were. In Example 1, observing a ruby ball provides severe evidence for H_R, even though your test has low power, because with probability one, you would not have pulled a scarlet ball were H_R false (and H_B true). In short, epistemologists' investigation of *basis-relative* notions of sensitivity are analogous to investigation of *data-dependent* notions of power by philosophers of statistics.

Just as there are challenges to the necessity of a simple version of sensitivity for knowledge, there are also also challenges to its sufficiency. Harman [2008], for example, asks us to imagine that a dictator dies and a small news station reports the story immediately. Suppose Ian hears the report and thus believes (rightfully) that the dictator is dead. However, imagine that, prior to the dictator's death, the government planned a cover-up. So news stations had been instructed not to report the dictator's death (whenever it happens). Consequently, the small news station issues a retraction minutes later, and all other news outlets continue to report that the dictator is alive. Ian, ignorant of the other reports, continues to believe the dictator is dead.

Here, Ian's belief seems to be sensitive: if the dictator were alive, he would not believe the dictator to be dead. Yet many think, intuitively, that Ian does not know that the dictator is dead. If he knew of the other news reports, he would believe that the dictator is alive and that the small news station had erred. Thus, true, sensitive belief is not sufficient for knowledge. An analogous case, again modeled on an example due to Berger and Wolpert [1988, p. 8, Example 4a], shows that a hypothesis passing a highly powered test might nonetheless be poorly supported by the evidence.

[15] For simplicity, we ignore the differences (if any) between Nozick's "methods," Sosa's "bases," etc.

Example 2 Imagine Ace Inc. makes two types of urns, both of which contain 100 balls. R-urns contain 1 ruby ball, 98 scarlet balls, and 1 black ball; B-urns contain 1 ruby ball, 0 scarlet balls, and 99 black balls.

Urn type	Ruby	Scarlet	Black
R	1	98	1
B	1	0	99

Table 2: Colors of the balls in the Ace urns.

You find an Ace urn, but you don't know which type. So you decide to run a Neyman-Pearson test on the hypothesis H_R "The urn is R-Type." You will draw one ball from the urn. If you draw a black ball, you'll reject H_R for H_B, "The urn is B-Type." Otherwise, you'll accept it.

Your test has high power: there's a 1% chance you will accept H_R if H_B is true. Yet, observing a ruby ball does not provide strong evidence for H_R, hence against H_B, because both R-urns and B-urns have exactly one ruby ball. So, having high power cannot be a sufficient condition for a test to yield strong evidence.

The problem with power, and with sensitivity, is that they are essentially vacuous when an observation is improbable *no matter which hypothesis is true*. In Example 2, it is highly and equally improbable that one will observe a ruby ball, regardless of whether H_R or H_B is true. So even though the proposed test of H_R has a high power against H_B, observing a ruby ball does not provide good evidence that H_R is true, and H_B false. And in Harman's case, the chances of Ian learning that the dictator is dead are virtually zero, regardless of whether the dictator is in fact dead. So although Ian's belief is sensitive because there is no other nearby world in which he believes the dictator is dead, it is not thereby knowledge.

This problem extends to basis-relative sensitivity and data-dependent severity. But, Nozick does not claim that sensitivity is sufficient for knowledge, nor do Mayo and Spanos claim that severity is sufficient for strong evidence for a hypothesis. The latter also require *fit*—that the evidence has a mediocre or large p-value given the hypothesis—while the former also requires an additional condition often called *adherence*, which states that if an agent knows φ, then in nearby worlds in which φ is true, she believes φ. We will leave others to summarize how significance levels (the non-data dependent version of p-values) and non-basis relative versions of adherence have been subjected to analogous counterexamples in epistemology and statistics, respectively. Instead, we discuss an objection to the necessity of basis-relative versions of adherence, due to Kripke [2011, p. 178], extending it to Mayo and Spanos's notion of fit. We then briefly argue that the objection reveals a strength, not a weakness of the requirement.

Imagine a box containing two slits through which a photon might pass. Suppose there is a detector plate behind the right slit but not the left one. Mary will believe a photon was emitted if and only if the detector is activated; otherwise, she will will suspend judgment. Finally, suppose that a photon passes through the right slit, activating the detector. Intuitively, Mary's belief counts as knowledge. But according to Nozick's theory, Mary does not know a photon was emitted because in nearby worlds in which the photon passes through the left slit, she does not believe a photon has been emitted (as she suspends judgment). The example shows that, although Mary got lucky confirming evidence, her belief still counts knowledge because she would not have received the same confirming evidence if her belief (that a photon was emitted) were false.

Example 1 above illustrates an analogous problem for the thesis that good evidence for an hypothesis must fit the hypothesis in Mayo and Spanos's sense, i.e., that in order for data to count as evidence for the null hypothesis, its probability (or that of data at least as extreme) should be high (or at least not low) if the null hypothesis is true. Drawing a ruby ball in Example 1 *intuitively* provides good evidence for H_R, yet one is unlikely to observe a ruby ball if H_R is true (and so one is unlikely to believe H_R on the basis of observing a ruby ball even if H_R is true).

We deny that intuition. Observing a ruby ball is *by itself* bad evidence for H_R. It is, of course, good evidence *against* H_B. So if one already had good evidence E for the disjunctive hypothesis $H_R \vee H_B$, then the *conjunction* of E and one's observation of a ruby ball might constitute good evidence for H_R. But a ruby ball alone does not. We conclude that Mayo and Spanos's theory delivers the correct verdict that a ruby ball, itself, is bad evidence for the hypothesis H_R.

Some might object that our defense of Mayo and Spanos's theory is a bit of a cop out, but we think it reveals a deep strength of their theory and a limitation of other approaches to measuring statistical evidence. Suppose now that you draw three balls with replacement from the unknown urn in Example 1. Imagine that, each time, you draw a ruby ball. That data might reasonably lead you to doubt your *background assumption* that the urn is Type R or B. Although your evidence still definitively rules out the hypothesis H_B, you have still witnessed an incredibly improbable outcome if the hypothesis H_R is true. So drawing three ruby balls seems to be good evidence against the disjunctive hypothesis $H_R \vee H_B$.[16]

An analogous response applies to Kripke's purported counterexample. If Mary's detector is activated every time that a photon *might* have been emitted, then Mary could reasonably become convinced that her detector is broken, or that something about the physical apparatus prevents photons from passing to one side. And that might reasonably undermine her claim to knowledge that any particular photon was emitted.

These responses illustrate how classical statistical testing avoids van Fraassen's "bad lot" objection [Van Fraassen, 1989, pp. 142–3], which, though originally aimed at inference to the best explanation, applies equally to any account of inference or evidence that is merely comparative. Such theories, such as likelihoodism and some forms of Bayesianism, claim that data only provide evidence for one hypothesis over another, rather than for or against a single hypothesis. The objection is that the best of bad lot can be no good at all: just because an hypothesis fares better than others under consideration needn't entail that it is should be believed, as sufficiently strong evidence ought compel. Requiring that the data fit an hypothesis to support it evidentially means that when all the statistical hypotheses under consideration fail to fit the data, they should all be rejected [Gelman and Shalizi, 2013]. This process of checking statistical modeling assumptions [Mayo and Spanos, 2004] can then in turn spur generation and consideration of new hypotheses.

3 Conclusions and Future Research

Although classical hypothesis testing was developed to answer concrete questions in the agricultural and social sciences whereas modal theories of knowledge were designed as solutions to abstract concerns about external world skepticism, the two share considerable theoretical structure. Both employ subjunctive conditionals relating, respectively, hypotheses and possible worlds with data and bases for belief, as logical conditions for evidence and knowledge. We reviewed a few counterexamples against some of these conditions in section 2, showing how the statistical and epistemology liter-

[16] The careful reader might note that this line of reasoning can also be used to defend the thesis that pre-sample power is necessary for evidence against the objection in Example 1. A small modification to that example (by increasing the number of ruby balls in Urn R to one million and leaving all other things the same), however, illustrates the same problem for power and provides further reason to think Mayo and Spanos's notion of "fit" is tracking intuitions.

atures have paralleled each other. Surely further examples from each could be adapted to bear on the other. For example, classical hypothesis testing requires an extremeness relation on possible data sets that bears some resemblance to the similarity relations on possible worlds used in the semantics for modal conditions on knowledge: the more extreme data are, the less probable they are in, with respect to some way of aggregating the data. Does this mean such data are less relevantly similar to the most probable data? What determines the best way to aggregate the data, i.e., the selection of a test statistic to summarize the data and whose values are the relata of the extremeness relation?

Another set of tasks we have not completed includes evaluating how well classical hypothesis testing fares in accomplishing typical goals for philosophical theories of evidence. These include specifying the ontology and logical structure of evidence, clarifying its role in justifying belief, constraining rationality, guiding inquiry towards the truth, and arbitrating these tasks objectively, publicly, and intersubjectively. Section 1 already makes some progress on the logical structure of evidence in classical hypothesis testing, and the introductory section mentioned that it is largely quietist about ontology. This makes sense if the motivation for the ontology of evidence arises from epistemologists' general assumption that evidence is first-person, thus must interface with the philosophy of mind. In contrast, evidence in classical statistics is third-personal. Although statisticians and scientists agree that it should rationally guide belief and decisions, that guidance itself is not a part of the theory of testing. Providing such guidance could be a philosophical project worthy of pursuit.

One of the prima facie advantages of third-personal evidence is its natural connection with objectivity. Because there are many sorts of objectivity [Douglas, 2009], it would be worthwhile to investigate whether classical statistical testing provides the sort that is epistemologically desirable—see, e.g., Reiss and Sprenger [2017, §4.2.2] for a dissent. In contrast with Bayesian statistics, for example, no use of utility functions or subjective prior probabilities over the space of hypotheses is needed, although (as alluded above) the choice of test statistic raises questions about what Douglas calls "procedural objectivity." Connected with objectivity, of course, is the sense in which evidence is supposed to be indicative of the truth. Here, further analysis of the significance of famous limiting theorems, such as the law of large numbers and the central limit theorem, deserve more philosophical attention.

References

Kelly Becker and Tim Black. *The sensitivity principle in epistemology.* Cambridge University Press, 2012.

James O. Berger and Robert L. Wolpert. *The likelihood principle.* Institute of Mathematical Statistics, Hayward, CA, 2nd edition, 1988.

George Casella and Roger L. Berger. *Statistical Inference.* Duxbury Press, Belmont, CA, 2nd edition, 2001.

David R. Cox. Some problems connected with statistical inference. *Annals of Mathematical Statistics*, 29:357–372, 1958.

Heather Douglas. *Science, policy, and the value-free ideal.* University of Pittsburgh Press, 2009.

Fred Dretske. Conclusive reasons. *Australasian Journal of Philosophy*, 49 (1):1–22, 1971.

John Earman. *Bayes or Bust?* MIT Press, 1992.

Samuel Fletcher. The logic of severe testing. *Unpublished Manuscript.*

Andrew Gelman and Cosma Shalizi. Philosophy and the practice of bayesian statistics (with discussion). *British Journal of Mathematical and Statistical Psychology*, 66:5–64, 2013.

Gerd Gigerenzer, Zeno Swijtink, Theodore Porter, Lorraine Daston, Jogn Beatty, and Lorenz Krüger. *The empire of chance: How probability changed science and everyday life.* Cambridge University Press, 1989.

Gilbert Harman. Thoughts: Selections. In Ernest Sosa, Jaegwon Kim, and Matthew McGrath, editors, *Epistemology: An Anthology*, pages 194–206. Blackwell Publishing, 2008.

John M. Hoenig and Dennis M. Heisey. The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):19–24, 2001.

Colin Howson and Peter Urbach. *Scientific reasoning: the Bayesian approach.* Open Court, Chicago, 3rd edition, 2005.

Thomas Kelly. Evidence. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016. URL `https://plato.stanford.edu/archives/win2016/entries/evidence/`.

Saul A. Kripke. *Philosophical troubles: collected papers*, volume 1. Oxford University Press, 2011.

E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, New York, 3rd edition, 2010.

Dennis V. Lindley. The estimation of many parameters. In V. Godambe and D. Sprott, editors, *Foundations of statistical inference*, pages 435–447. Holt, Rinehart and Winston of Canada, Toronto, 1971.

Deborah G. Mayo. *Error and the growth of experimental knowledge*. University of Chicago Press, 1996.

Deborah G. Mayo. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press, 2018.

Deborah G. Mayo and Aris Spanos. Methodology in practice: Statistical misspecification testing. *Philosophy of Science*, 71:1007–1025, 2004.

Deborah G. Mayo and Aris Spanos. Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science*, 57:323–357, 2006.

Deborah G. Mayo and Aris Spanos. Error statistics. In M. Forster and Prasanta S. Bandyopadhyay, editors, *Philosophy of statistics*, number 7 in Handbook of the Philosophy of Science, pages 153–198. Elsevier, 2011.

Conor Mayo-Wilson. Epistemic Closure in Science. *Philosophical Review*, January 2018.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

Robert Nozick. *Philosophical explanations*. Harvard University Press, 1981.

Julian Reiss and Jan Sprenger. Scientific Objectivity. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*.

Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017. URL https://plato.stanford.edu/archives/spr2017/entries/scientific-objectivity/.

Gail Stine. Skepticism, Relevant Alternatives, and Deductive Closure. In Ernest Sosa, Jaegwon Kim, and Matthew McGrath, editors, *Epistemology: An Anthology*, pages 247–255. Blackwell Publishing, 2008.

Aleksandr B. Tsybakov. *Introduction to nonparametric estimation.* Springer, 2009.

B.C. Van Fraassen. *Laws and Symmetry.* Oxford University Press, USA, 1989.

Larry Wasserman. *All of statistics: a concise course in statistical inference.* Springer, 2004.